Session 5A: Research Infrastructures: ensuring trust and quality of data

Darren Bell

Repository Architect – UK Data Service

ICRI 2018 Wien

13 Sep 2018

UK Data Service

Infrastructure

- There is a plethora of standards, communities, regulators, actors, projects and frameworks. And that's just social sciences.

- Trust and Quality often based on an accreditation model rather than intrinsically evident from the data artefacts themselves.

- http://seriss.eu Any infrastructure consists of actors, agents, roles, processes, objects and tools.

- Real "Trust" (and assurance of Quality) require an understanding of what each of these entities actually are and what events have occurred in relation to them.

Infrastructure (2)

- Big Data and HPC tech is now changing the game.
- There is a risk that the gap between HPC infrastructures and traditional “boutique” repositories will widen.
- We are now in a world where the traditional Repository has to move on from a bibliographic perspective to a data science perspective if it is to stay relevant.
- This means promoting the importance of data stewardship and the professionalization of roles related to data authority and ownership.

The tension between Open Science and Privacy

- Within the social sciences certainly, most useful data is personal data
- Many techniques to anonymise and protect this sensitive data but can throw up many barriers to usage and sharing of data.
- To do FAIR properly, we must look to promote more machine-actionable access and rights models:
  W3C ODRL https://www.w3.org/TR/odrl-model/
  SPECIAL https://www.specialprivacy.eu/
- Provenance chains contribute directly to trust but realistically these can no longer be generated solely by humans.
  See PROV https://www.w3.org/TR/prov-overview/

Standards

- **Standardisation** (and consolidation) of meta/data object models that enable better interoperability, better sharing and robust versioning of data.

- At the UKDS, we have witnessed with dismay one new “universal portal” after another creating a new metadata standard.

- Championing a small number of taxonomies and metadata standards across domains, particularly in areas of QA operations and privacy.

- Will lead to the realistic possibility of citizen access to their own data and information about its usage.

HM Treasury – the Economic Value of Data

Discussion Paper - August 2018

“Provided data sharing is safe, ethical and compliant with data protection laws, the government wants to make it easier for firms to share useful data, to support growth and innovation. A lack of effective data sharing would only serve to concentrate power within the hands of a few large businesses, and stifle the innovation, quality, and value for money that that arises normally in markets through effective competition and disruption.”

…

…

“the UK Anonymization Network provides support and information to businesses looking to undertake anonymization of personal data. However, these resources are not always widely known, and many businesses are still nervous about the implications and risks associated with anonymization. There is therefore a continuing role for public sector bodies and other key stakeholders in setting out principles for the safe, secure, and effective anonymization of personal data, in order to enable effective data sharing.”

New models

- Accelerated adoption of **new data paradigms** should be promoted, such as linked open data and NoSQL:

  (1) The harvester/provider model cannot scale. More "peer to peer" and fewer "hub and spoke" connections between RIs.
  (2) Models to enable domain-agnostic data that breaks down the traditional demarcation lines between disciplines.

- Smart Meter Research Portal project at UKDS is a deliberate attempt to do interdisciplinary data end-to-end with energy, environmental and social science data, based on big data and semantic tech.

- We need to fight "portal sprawl" and look to distributed models for data analytics and dissemination.

Data creators

- FAIR is a good framework for setting expectations related to the data consumer BUT

- New supply/ingest models are required to handle the scale and periodicity of **new and novel forms of data**, demanding a full lifecycle approach to RIs.

- We don’t just need policies for depositors, we need toolchains.

Summary

- Trust should be an emergent property of privacy by design, not just an assertion based on accreditation
- Quality should be an emergent property of provenance by design
- “Design” implies formal object models and commonly accepted machine-actionable standards
- Rather than building mega “hubs” that have the whiff of five year plans, Government and academia should promote distributed models based on semantic web APIs and adoption of interoperability standards.
- We should encourage “fail-fast” bilateral experiments between different disciplines and countries (beyond Europe) and iteratively build a real Web 3.0